Psychology Science, Volume 45, 2003 (1), p. 106-110

On artificial results due to using factor analysis for dichotomous variables

Klaus D. Kubinger[1]

Summary

This paper serves to remind the reader, that factor analysis in the case of dichotomous variables will often lead to artificial factors. In other words, the factors correspond primarily to certain levels of item difficulty. A numerical example will be given in order to illustrate this. It is argued that, for instance, factoring tetrachoric correlations instead of conventionally used Pearson correlations would lead to more content valid results.

Key words: Factor analysis, dichotomous variable, artificial factor, tetrachoric correlation

Introduction

During the last forty years factor analysis has become one of the most used methodical approaches within psychology, and hundreds of psychological tests have been developed based on factor analysis. Yet, the use of dichotomous variables is problematic. As a matter of fact, factor analysis applied to dichotomous variables leads to artificial results. This paper will, in the following, try to remind the reader of this and to illustrate it with a numerical example. Doing so seems necessary, because psychological test constructors still use, almost obligatorily, a factor analysis for items that are scored dichotomously. This is true for instance for *Jackson*s world-wide renown test-battery PRF *(Personality Research Form).*

The paper will not deal with the problem of using ordinal scaled variables for factor analysis. However the reader should bear in mind that a factor analysis does require interval-scaled variables because of the use of the *Pearson* correlation; and that psychological data are very often simply ratings which are not likely to establish equal distances (see for instance the group of studies concerning *Osgood*s semantic differiential).

[1] Univ.-Prof. Dr. Mag. Klaus D. Kubinger, Division for Assessment and Applied Psychometrics, Department of Psychology, University of Vienna, Liebiggasse 5, A-1010 Wien, Europe, E-mail: klaus.kubinger@univie.ac.at

The problem

Now let us turn to dichotomous variables: a prudent user of a factor analysis would have realized, for instance, the hints already Torgerson (1958, p. 331) has given: "Factor-analysis procedures would probably give a good selection [of dichotomously scored items in order to increase a test's scalability] as long as the coefficients that are factored do not depend unduly on item marginals. Where chance success is not an element, tetrachoric correlations or the ratio of the phi coefficient to the maximum possible phi for the item marginals seem appropriate." However, according to Fergerson (1941), Smith (1950), Stouffer, Guttman, Suchman and Lazarsfeld (1950), and Guttman (1955) the evidence already existed some years ago that by using dichotomous items – as a consequence of which the *Pearson* correlation reduces the ϕ-coefficient – factor analyses will most likely result in as many factors as there are items with different item difficulties (the latter correspond to the item marginals). Sixtl (1969) also intensively elaborated on this matter.

To remind the reader so far!

We think that, as time goes by, potential users of a factor analysis will still require certain illustrations. Let us consider a numerical example for this.

Figure 1 gives initially the data of n = 100 subjects as they mastered (+) or failed (–) both the items 1 and 2 of any item pool (see example 1 in the upper left hand table). The ϕ-coefficient amounts to .40. Of course, this correlation is rather small and the explained variance is just 16%. The application of a factor analysis would most likely not lead to high factor loadings for both the items with regard to the same factor. However, this example is conceived as having different item marginals, in other words, the difficulty of the items is different. While item 1 is mastered by only 20 out of 100 subjects, item 2 is mastered by 50 out of 100. Therefore, even if the item association would have been ideal, given the marginals, the ϕ-coefficient never reaches 1. The maximum possible ϕ-coefficient ϕ_{max} equals .50 here (see ideal association in the upper right hand table)! Taking this ϕ_{max} into consideration, the ϕ-coefficient .40 no longer looks so small. Some moderate or high association is therefore obvious. However the statistical coefficient does not reflect this. Only if both the item marginals were equal, could the ϕ-coefficient possibly be 1 (see example 2 in the lower left hand table). This proves the statement made above that only those items which have (exactly) the same item difficulties are likely to create a factor – one should, however, take into account that, of course, two items may have no association at all even though they have the same difficulties. For an even better illustration of ϕ-coefficient's dependency on the equality of the item marginals in Figure 1 a case will finally be given of item marginals which are more similar (see example 3 in the lower right hand table): Instead of .50 the maximum possible ϕ-coefficient ϕ_{max} then amounts to .65.

Numerical example 1

		ITEM 1 +	ITEM 1 -	
ITEM 2	+	18	32	50
	-	2	48	50
		20	80	

$\phi = .40$

Ideal association, given the marginals

		ITEM 1 +	ITEM 1 -	
ITEM 2	+	20	30	50
	-	0	50	50
		20	80	

$\phi_{max} = .50$

Example 2: Ideal association

		ITEM 1 +	ITEM 1 -	
ITEM 2	+	20	0	20
	-	0	80	80
		20	80	

$\phi_{max} = 1.00$

Example 3: Ideal association

		ITEM 1 +	ITEM 1 -	
ITEM 2	+	30	20	50
	-	0	50	50
		30	70	

$\phi_{max} = .65$

Figure 1:
A numerical example and some ideal item associations. The maximum possible ϕ-coefficient differs in dependence of the item marginals. It only then reaches 1 if the marginals of two items are the same.

Consequences

As indicated above by Torgerson, the problem might possibly be solved by using a tetrachoric correlation instead of a *Pearson* correlation and ϕ-coefficient, respectively. His alternative suggestion, that being to use the ratio of the ϕ-coefficient to ϕ_{max}, does not seem to be a proper means, because although this ratio frequently applies in psychological research it is nevertheless merely a descriptive statistic; therefore nothing is really known about its underlying distribution and so, no significance test exists. The opposite is true with regard to the tetrachoric correlation (see for instance Kubinger, 1990/1993).

The formula for the tetrachoric correlation results if the *Pearson* correlation is applied to dichotomous variables, however these variables are supposed to be normally distributed from the start (Pearson, 1907). Although Kendall and Stuart (1979) give a more detailed formalisation for calculating the tetrachoric correlation, the equation system given by Dixon (1985) does it here:

$$\Phi(u_1) = (a + c)/n$$
$$\Phi(u_1) = (a + b)/n$$

$$\int_{-\infty}^{u_1} \int_{-\infty}^{u_2} f(x, y, r_{tet})\, dxdy = a/n,$$

$\Phi(u)$ is the distribution function of a standard normal distribution at u, with $f(x, y, r_{tet})$ being the density of a standardized bivariate normal distribution at (x, y) given the correlation

r_{tet}; and a, b, c, (and d) being the frequencies in the contingency table (e.g. a for the counts of cell ++, b for – +, c for + –, and d for – –). However, for even more simplicity there is also a formula of approximation:

$$r_{tet} = \cos \{180^o / [1 + \sqrt{(bc/ad)}]\}$$

– which should be chosen in the case of a, b, c, and d so that a and/or d are not zero.

Obviously, r_{tet} results in 1 if b times c is zero; this always happens if an ideal association, for instance according to Figure 1, occurs. It does not matter whether the item marginals are equal or not. In the numerical example in Figure 1 r_{tet} is approximated as (–).7825.

Unfortunately, the tetrachoric correlation is not available in the standard software package SPSS - neither solely for the purpose of calculations by using "cross tables", nor by using the subroutine "factor analysis". So far our considerations here run the risk of being useless. For this a special SPSS syntax software is currently in preparation which will soon be available to everyone on our homepage: www.univie.ac.at/psychologie/diagnostic.

Conclusion

For psychologists who are acquainted with structural equation models (e.g. LISREL, "Linear Structural Relationships") approaches for factoring dichotomous variables that are based on these models may be recommended (c.f. in particular Muthén & Christoffersson, 1981, Muthén, 1984, 1989). Knol and Berger (1991) and Parry and McArdle (1991) did in fact compare such approaches with a factor analysis using the tetrachoric correlation – and basically found that factoring tetrachoric correlations worked as well. However the aim of this paper is not to disclose any conditions when concurrent algorithms are superior; the purpose of this paper is simply to prevent psychologists who do not have thorough knowledge of statistics from using conventional (i.e. SPSS-default) factor analysis for dichotomous variables.

References

1. Dixon, W.J. (Ed.)(1985). BMDP statistical software. Manual. Berkeley: Univ. of California Press.
2. Ferguson, L.W. (1941). The factorial interpretation of test difficulty. Psychometrika, 6, 323-329.
3. Guttman, L.A. (1955). A generalized simplex for factor analysis. Psychometrika, 20, 173-192.
4. Kendall, M. & Stuart, A. (1979). The advanced theory of statistics, 2. London: Griffin.
5. Knol, D.L. & Berger, M.P. (1991). Empirical comparison between factor analysis and multidimensional item response models. Multivariate Behavioral Research, 26, 457-477.
6. Kubinger, K.D. (1990/1993). Übersicht und Interpretation der verschiedenen Assoziationsmaße [Review of pertinent measures of association]. Psychologische Beiträge [later: Psychology Science], 32, 290-346.
7. Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. Psychometrika, 46, 407-419.
8. Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 407-419.
9. Muthén, B. (1989). Dichotomous factor analysis of symptom data. Sociological Methods and Research, 18, 19-65.

10. Pearson, K. (1907). On further methods of determining correlation. Draper's Company Memoirs, Biometric Series IV.
11. Parry, C.D. & McArdle, J.J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. Applied Psychological Measurement, 15, 35-46.
12. Sixtl, F. (1969). Difficulty, complexity, reliability: Critical observation concerning the theoretical conception of psychometric tests. In K.H. Ingenkamp (Ed.), Developments in educational testing, 1 (pp. 414-422). London: Univ. of London Press.
13. Smith, R.G.Jr. (1950). A factorial study of attitude toward the negro. Doktordissertat. Univ. Illinois.
14. Stouffer, S.A., Guttman, L., Suchman, E.A. & Lazarsfeld, P.E. (1950). Measurement and prediction. Princeton: Princeton Univ. Press.
15. Torgerson, W.S. (1958). Theory and methods of scaling. New York: Wiley.
